NOT TO BE FURTHER COPIED OR REPRODUCED

(* BATCHES *)

7

Effective Comparison

(well-chosen expression)

		Exh.		Page
	Introduction 7-1			
		1	a	7-3
		2		7-5
7A.	Use of central values 7-1	3		7-7
	- a precipitation example, 7-2 to 7-6	4		7-8
	- other examples, 7-6 to 7-9			
			b	
7B.	Comparisons are everywhere 7-10		c	
		5		7-12
		6		7-13
7C.	Making comparisons more useful 7-11	7		7-15
	- a plankton example, 7-11 to 7-14	8		7-17/18
	- back to warp breaks, 7-14 to 7-16	9		7-19
			d	
7D.	Choosing expression 7-16	10		7-21
	- some artificial examples, 7-20	11		7-22
	- warp breaks again, 7-20 to 7-23	12		7-24
		13		7-25
			e	
7E.	The meaning of comparison 7-23		f	
		14		7-28
7F.	Rough adjustments and rough residuals 7-27	15		7-29
	- plankton again, 7-27 to 7-30	16		7-31
	- combined approaches again, 7-30 to 7-32			
		17	g	7-33
7G.	The pursuit of symmetry takes much data 7-32	18		7-36
	- back to volcanoes, 7-32 to 7-35	19		
	- precipitation again, 7-35 to 7-38			
	- a comment, 7-38 to 7-39		h	
	**********	20	k	7-44/45
7H	How far have we come? 7-40	21	l	7-47
	**********	22		7-49
		23		7-50
7K.	Effects of re-expression (background) 7-42	24		7-51
		25		7-52
7L.	A psychological example (optional) 7-46	26	p	7-55
	**********	27		7-57
		28		7-58
7P.	Problems for chapter 7 7-54	29		7-59
		30		7-60
		31		7-61
		32		7-62
		33		7-63

7

Effective Comparison
(well-chosen expression)

As the text of this chapter will convince us more strongly than we see now, comparison is of central importance in all sorts of data analysis. Doing it well -- especially by avoiding unnecessary interferences -- is thus very important.

Part of doing well is to choose effective techniques. We have given considerable attention to this.

Another part is to use them well. We need to give more attention to this. Hence this chapter.

7A. Use of central values

Perhaps the simplest form of comparison -- one that is sometimes adequate -- is comparison in terms of a single number. When we are comparing levels, this number will almost certainly be a central value. So far, we have focused on the trimean (supplemented by the midmean) as a useful standard technique. What we need to do now, in order to clear up the use of central values for comparison, is to:

- understand when we might want to use some other central value.
- consider illustrations of how well central values summarize batches in comparisons.

Sometimes the values of a batch are only important as a total. The losses sustained by one company under

insurance policies of a certain kind and the results of one individual's flutters in Wall Street are business examples. Most golf tournaments depend on a total count of strokes; a catastrophic hole ruins a player. Similarly, a pitcher's earned run average includes his bad days. In such circumstances, we ought to use only the mean as a central value (though we are likely to want, even in such cases, to figure residuals from a middle mean).

In other circumstances, where, conversely, we wish to minimize the effect of stray observations, which may be rather more frequent than we would like, the trimean and the midmean may not be resistant enough for us. At one time, the median would have seemed to be indicated. Today, the skipping trimean -- or skipping midmean -- seems likely to meet our needs even more satisfactorily.

Such special situations aside, the trimean -- or the midmean -- is very likely to be a good choice as a central value.

(* a precipitation example *)

An interesting example is the annual precipitation (rain plus snow) in New York City for a set of six-year periods starting every 10 years from 1890 on. Exhibit 1 gives the data. Looked at as a whole, it merely looks like another table of figures.

exhibit 1 of chapter 7

Precipitation* in New York City (to nearest inch)
in years ending with 0,1,2,3,4 and 5

	189-	190-	191-	192-	193-	194-	195-
-0	52	42	36	49	35	45	45
-1	41	47	40	34	36	36	47
-2	39	47	38	43	39	50	46
-3	53	49	44	37	50	40	38
-4	44	42	34	38	45	52	43
-5	36	44	41	37	33	46	41

*Rain plus rain equivalent of snow.

Available from a variety of sources, including The World Almanac.

It would be helpful to look a little deeper into this array of numbers. Do there seem to be changes from decade to decade? If so, in what directions and in what amounts? Clearly it might help to look at a suitable central value for each six-year block and see whether the message -- if there be any -- is clearer.

Exhibit 2 shows the resulting comparisons. All of the kinds of central values we have used are shown, each in its own line. In this example, the mean tells the same story as the median, or the trimean:

- In the 1890's, 1900's, 1940's and 1950's, high typical annual rainfalls, near 44 inches,
- In the three intervening decades, lower typical rainfalls, near 38 inches.

There is now an understandable message -- 1910 to 1935 seems to have been a period of lowered rainfall. Whether or not this ought to be regarded as accidental is a confirmatory question, a question we have not tried to answer -- and will not try to answer here. However it should be regarded, its existence might well have been a warning to prepare for the "drought" of the 1960's. Our concern is that use of any one of the central values uncovers a message in this instance -- and that their use in other instances may also uncover messages.

In judging how well central values do this job, we ought to compare the simple numerical summary with its prime competitors:

exhibit 2 of chapter 7

Mean, median, and midmeans of six-year blocks of rainfall given in exhibit 1

	Initial Year of Block						
	1890	1900	1910	1920	1930	1940	1950
Mean	44.2	45.2	38.8	39.7	39.7	44.8	43.3
Median * **Midmean**	42.5	45.5	39.0	37.5	37.5	45.5	44.0
Trimean	44.0	45.0	38.8	38.8	38.8	45.2	43.8

* The midmean is the median, since there are no more than six values per batch.

- stem-and-leaf displays.
- schematic plots.

Exhibits 3 and 4 do just this for the New York City precipitation. In this instance, if we are really willing to look at numbers, the midmeans alone seem to tell us all the interesting things that the schematic plots do. (Clearly, this will not always be so.) The expanded version of the stem-and-leaf, if looked at very carefully, seems to tell us almost as much as the midmeans alone. In situations like this, the bare midmeans do a rather good job. Clearly there must be many places where they will serve us well.

(* other examples *)

If we turn back to the preceding chapter, we find a number of situations where single central values, specifically medians, do not tell the story that a broader comparison would. These examples are a little artificial, since they involve comparing an initial batch with what remains after the set-aside process has been applied. They are, however, also more convincing because we do know the true nature of the difference between what is being compared.

Exhibit 6 of chapter 6 dealt with the Cushny and Peebles data. The right-hand column gives a skipped analysis where it is clear that, though the median is exactly the same for initial and remaining batches, this equality tells us the least interesting fact about the comparison. Here the

exhibit 3 of chapter 7

Three kinds of summaries of New York City precipitation compared (third kind is exhibit 4)

A) STEM-and-LEAF COMPARISON -- unit 1 inch

	1890 -95	1900 -05	1910 -15	1920 -25	1930 -35	1940 -45	1950 -55
6							
5	23				0	02	
4	14	234779	014	39	5	056	13567
3	69		468	4778	3569	6	8
2							

B) SAME, OPENED UP -- unit 1 inch

5							
5	23				0	02	
4		779		9	5	56	567
4	14	224	014	3		0	13
3	69		68	778	569	6	8
3			4	4	3		

C) MIDMEANS -- unadorned

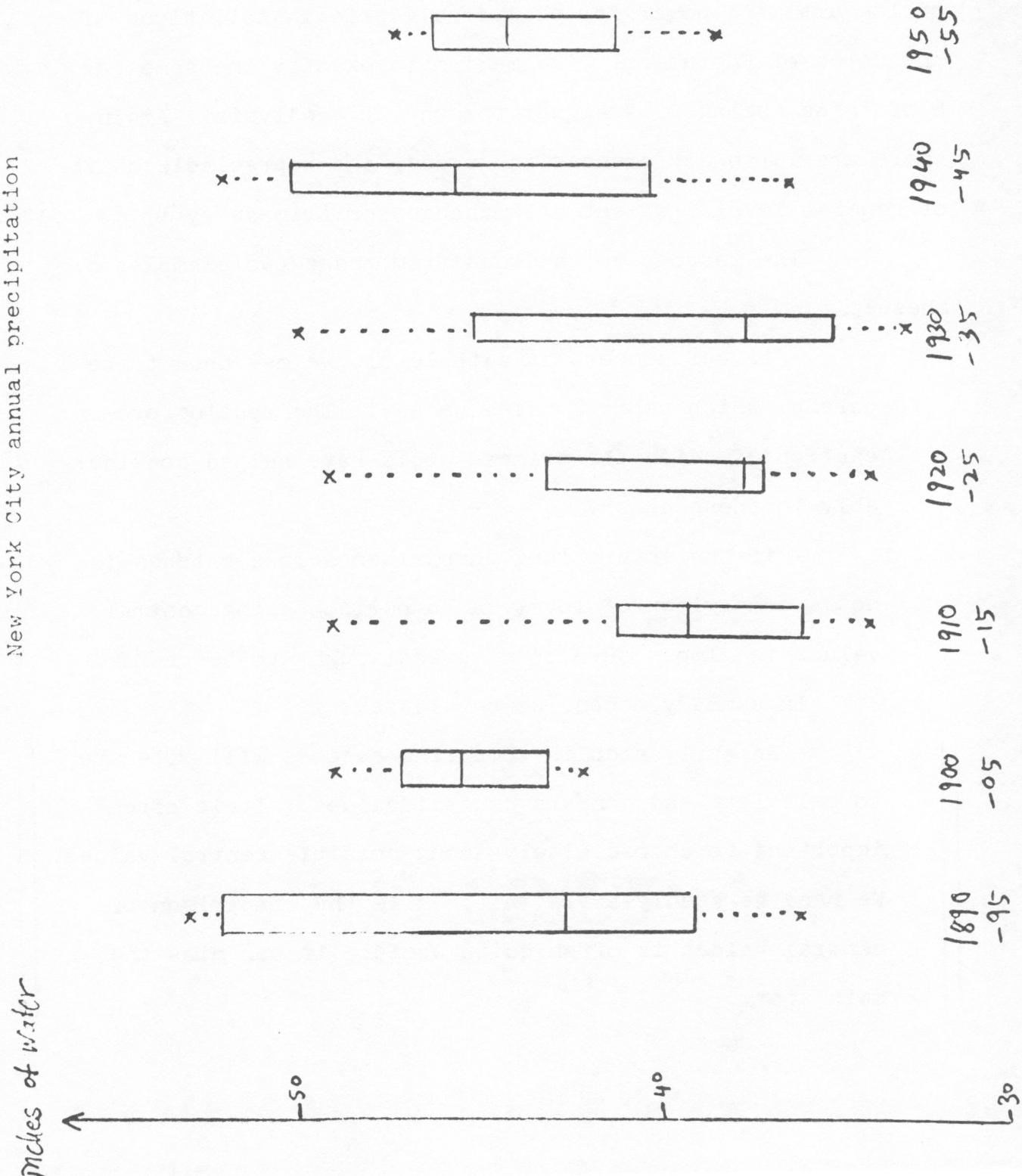
	42.5	45.5	39.0	37.5	37.5	45.5	44.0

D) SLIGHTLY GRAPHIC MIDMEANS

44:45		45.5				45.5	44.0
42:43	42.5						
40:41							
38:39			39.0				
36:37				37.5	37.5		

exhibit 4 of chapter 7

Third kind of summary for New York City precipitation



spread of the two batches differ by a large factor, and the set-aside batch runs somewhat smaller in value, as is reflected by the trimeans of 1.35 and 1.25.

Exhibit 9 of chapter 6 compares 7 different analyses of log ultimate power for 62 hydroelectric installations of the Corps of Engineers. The median is exactly the same for 6 of these, being 0.04 higher for the "U" analysis. Again, there are large differences in spread, and appreciable shifts of general level (reflected in changes of trimean by up to 0.12).

The lessons to be drawn from these two examples are these:

- if our concern is with level, we may need to be careful which central value we use. The routine precaution of using the trimean would have helped considerably in these cases.

- if the interesting comparison across batches is not a comparison of level, a comparison using central values is almost sure to miss what ought to be seen.

In summary, then, we should say:

- An early step in comparing batches will often be to calculate and compare central values. It is often important to choose wisely among possible central values. We need be keenly aware that, while the comparison of central values is often quite useful, it can miss the main story.

7B. Comparisons are everywhere

This chapter emphasizes comparison. We need to understand that not all comparisons are direct. Indeed, the importance of indirect comparisons far outweighs that of direct comparisons.

In a deep sense, almost all central values are for comparison. When we say that the average man in a newly discovered savage tribe is 6 feet 6 inches tall -- or 4 feet 4 inches -- the greatest impact of this figure is because the reader or listener compares it -- with other tribes, with his own country, with a mental foot rule, or with ideas of what heights are reasonable for Homo sapiens. (A report of 8 feet 8 inches would arouse great doubt!)

There are instances of measurement for supposedly higher purposes -- for insertion into formulas expressing the fundamental laws of physics, for example. But behind the practice of such measurement is much comparison, often chains of it, as when working scales are compared with secondary standards, secondary with primary standards, primary with national standards, and so on.

Summary values are usually for comparison, direct or indirect. What is good for comparison is good for almost any from of summarization. Thus the conclusions of this chapter apply very broadly.

7C. Making comparisons more useful

If we are not sure that central values will tell the story, we are likely to want to look at visual summaries. When we do, we want our schematic plots to tell us the story as dearly and as simply as they can.

When complications are inevitable, they are at least news, and may well be important. When complications are due to our own clumsiness, they are painful, confusing, perhaps dangerous, certainly wasteful and inefficient. This applies to comparing batches in particular, and to all forms of data analysis in general.

We always plan to express the data in a way that will help us as much as possible. This is as true in comparing batches as elsewhere.

In simple situations, we will usually clear up our perceptions when we make the batches to be compared more nearly equal in spread. There may be no way to do this, it may be possible but difficult, or it may be easy. When we can do it, with bearable effort, we usually gain by doing it. We get this gain whether we use visual summaries or numerical ones.

(* a plankton example *)

One way to seek more equal spread is to choose a better way to express our data. Exhibit 5 gives, in numbers, a simple example from Winsor and Clarke's (1940) analysis of the catch of plankton using different kinds of nets. Exhibit 6 shows schematic plots for the raw numbers,

exhibit 5 of chapter 7

Estimated number of three kinds of plankton caught in six hauls of each of two nets

A) RAW NUMBERS -- outside and detached points marked

	Kind I	Kind IV	Kind III
	1,050	15,500 ●	43,300 ●
	1,020	11,000	39,500 ○
H-	895	10,200	34,600
	845	9,830	32,800
	760	9,650	32,800
M-	620	9,250	31,400
	540	8,900	29,000
	537	8,600	28,800
H-	497	8,260	28,100
	470	7,900	27,800
	428	7,600	22,300
	387	6,060 ○	18,900 ●

B) SUMMARY VALUES

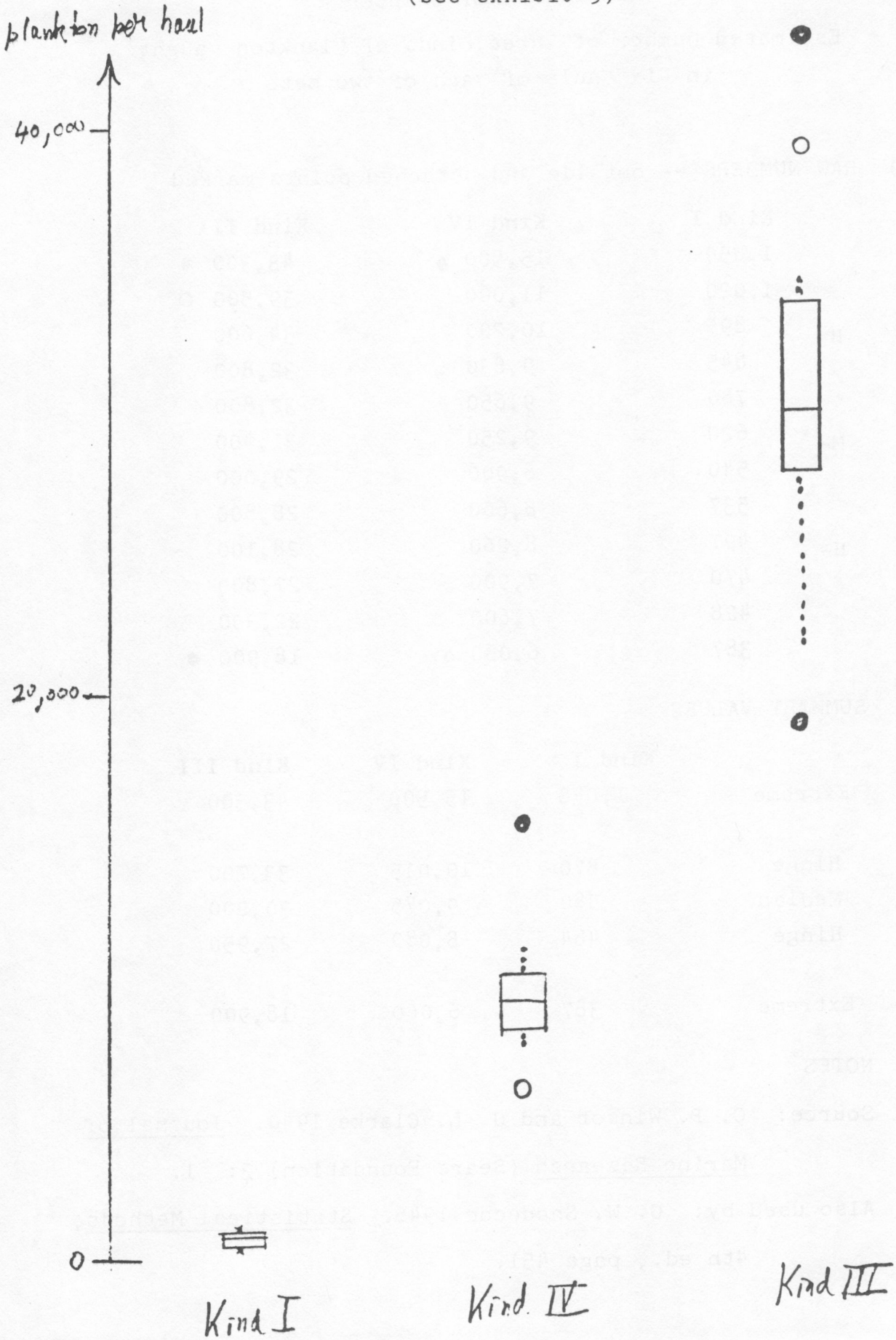
	Kind I	Kind IV	Kind III
Extreme	1,050	15,500	43,300
Hinge	870	10,015	33,700
Median	580	9,075	30,200
Hinge	484	8,080	27,950
Extreme	387	6,060	18,900

C) NOTES

Source: C. P. Winsor and G. L. Clarke 1940. Journal of Marine Research (Sears Foundation) 3: 1.

Also used by: G. W. Snedecor 1946. Statistical Methods, 4th ed., page 451.

exhibit 6 of chapter 7
Winsor and Clarke's estimated number of plankton per haul of net
(see exhibit 5)



which is not very clear. The difference in typical catch of the three kinds of plankton is all too obvious, but the great differences in spread tend to mask any detailed examination.

Especially since the natural change in catch is by such-and-such %, not by so-and-so many plankton, it seems natural to try the use of logs. When we do, we find the results in exhibit 7. Now our spreads are more nearly constant, and we get a fairly detailed idea of how the catches of the three kinds of plankton compare. In particular, apparent tendencies toward:

- one high and one low value for Kind IV,
- at least one low value for Kind III,

are reasonably evident, as is the rather clearer resolution of the difference between Kind I and Kind IV, in comparison with the difference between Kind IV and Kind III.

As this example has illustrated, proper choice of expression, by keeping spread from moving systematically as the central value varies from one batch to another, can be a real help in comparing batches of values.

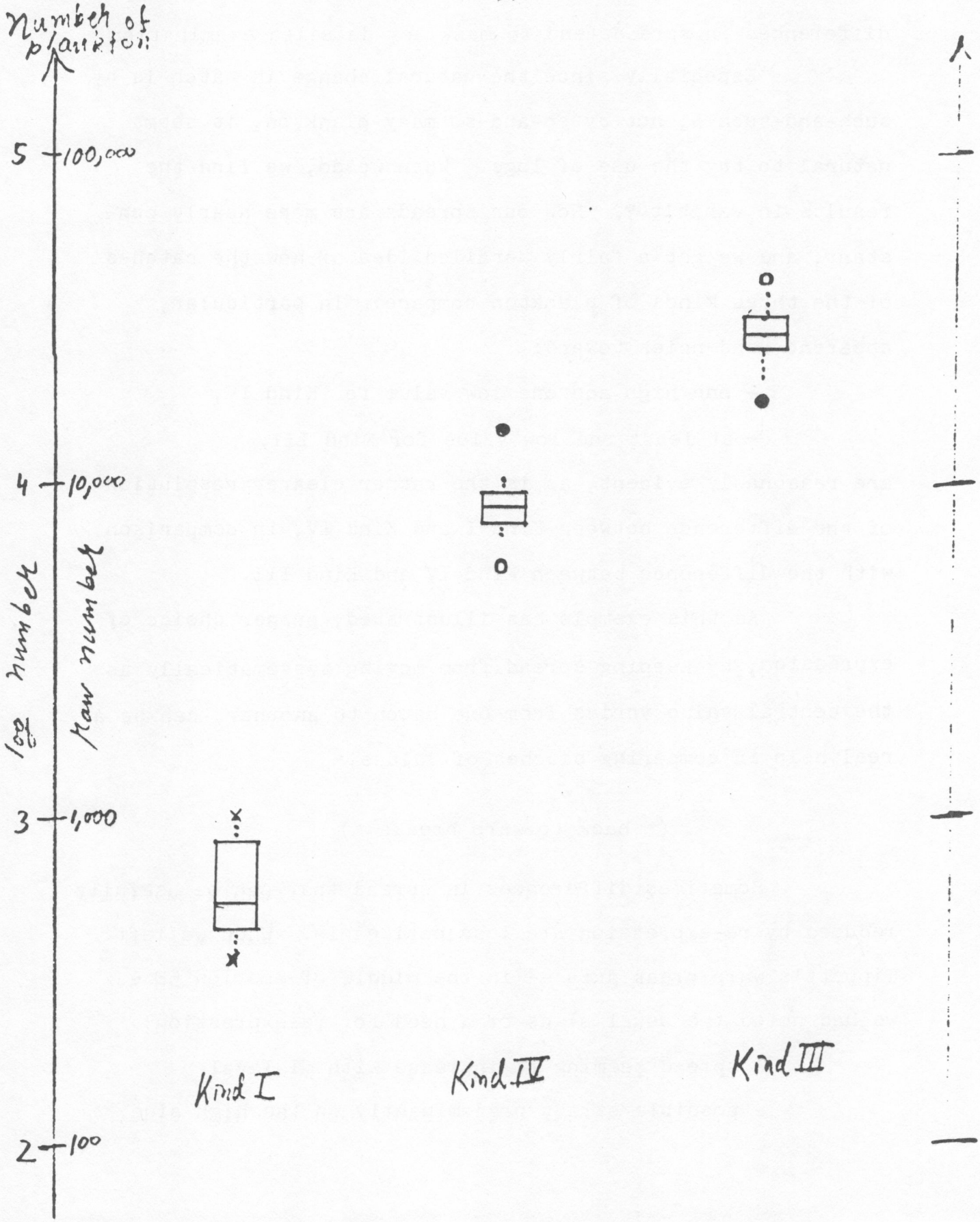
(* back to warp breaks *)

Sometimes differences in spread that can be usefully reduced by re-expression are less noticeable. When we left Tippett's warp break data -- in the middle of Section 5E -- we had noted the usual signs of a need for re-expression:

- spread seeming to increase with midlevel
- possible strays predominantly on the high side.

exhibit 7 of chapter 7

Winsor and Clarke's estimated number of plankton per haul of net (log scale)
(see exhibit 5)



(Spread decreasing with level, and strays toward the low side, would have suggested the infrequent direction of change of expression -- toward squares and cubes.)

If we are conservative, we may well try:

- square roots of counts as the manner of expression.
- plotting the deviations of these roots from their midmeans as a way to focus our attention on what happens within batches.

Exhibit 8 gives the numbers, exhibit 9 plots the picture. Rather clearly there is still a tendency for spread to increase with midmean. Presumably we should pass on to trying, as the next step stronger than roots, logs of counts. We shall leave trying this to the reader. (See problems below.)

7D. Choosing expression

In the plankton example, there were common sense reasons for using logs. A combination of common sense and the considerations discussed in the introduction to chapter 3 will lead us to good expressions in many instances.

In others, simple trial of the expressions that seem reasonable will show us what we want to do.

As we work with more and more data, we can gain more from a less trial-and-error procedure. There are three reasons for this:

- we may have enough data to show that some expression intermediate between those we might try is actually to be prefered. This does not happen very often.

exhibit 8 of chapter 7

Tippett's warp breaks in square roots

	BH	AM	AH	BM	BL	AL	(all)
A) SQUARE ROOTS of COUNTS, ORDERED -- detached and outside values marked							
	5.3	6.0	6.6 •	6.5	6.6	8.4	(6.57)
	4.9	5.9	6.0	6.2	6.4	8.2	(6.27)
	4.6	5.5	5.3	6.2	5.7	7.3	(5.77)
	4.5	5.4	5.1	5.4	5.4	7.2	(5.50)
	4.1	4.6	4.9	5.3	5.4	7.1	(5.23)
	4.0	4.2	4.6	5.1	5.2	5.5	(4.66)
	3.9	4.2	4.2	4.6	4.5	5.1	(4.40)
	3.9	4.1	3.9	4.4	4.4	5.1	(4.30)
	3.6	3.5	3.2	4.0	3.7	5.0	(3.83)
B) RANKED VALUES -- and MIDDLE MEANS							
end	5.3	6.0	6.6	6.5	6.6	8.4	(6.57)
next	4.9	5.9	6.0	6.2	6.4	8.2	(6.27)
hinge	4.6	5.5	5.3	6.2	5.7	7.3	(5.77)
median	4.1	4.60	4.9	5.3	5.4	7.1	(5.23)
hinge	3.9	4.2	4.2	4.6	4.5	5.1	(4.40)
next	3.9	4.2	3.9	4.4	4.4	5.1	(4.30)
end	3.6	3.5	3.2	4.0	3.7	5.0	(3.83)
trimean	4.18	4.72	4.82	5.35	5.25	6.65	
midmean	4.2	4.73	4.87	5.27	5.33	6.6	

(All) = Mean of values in first six columns.

(Cont'd)

exhibit 8 (Cont'd)

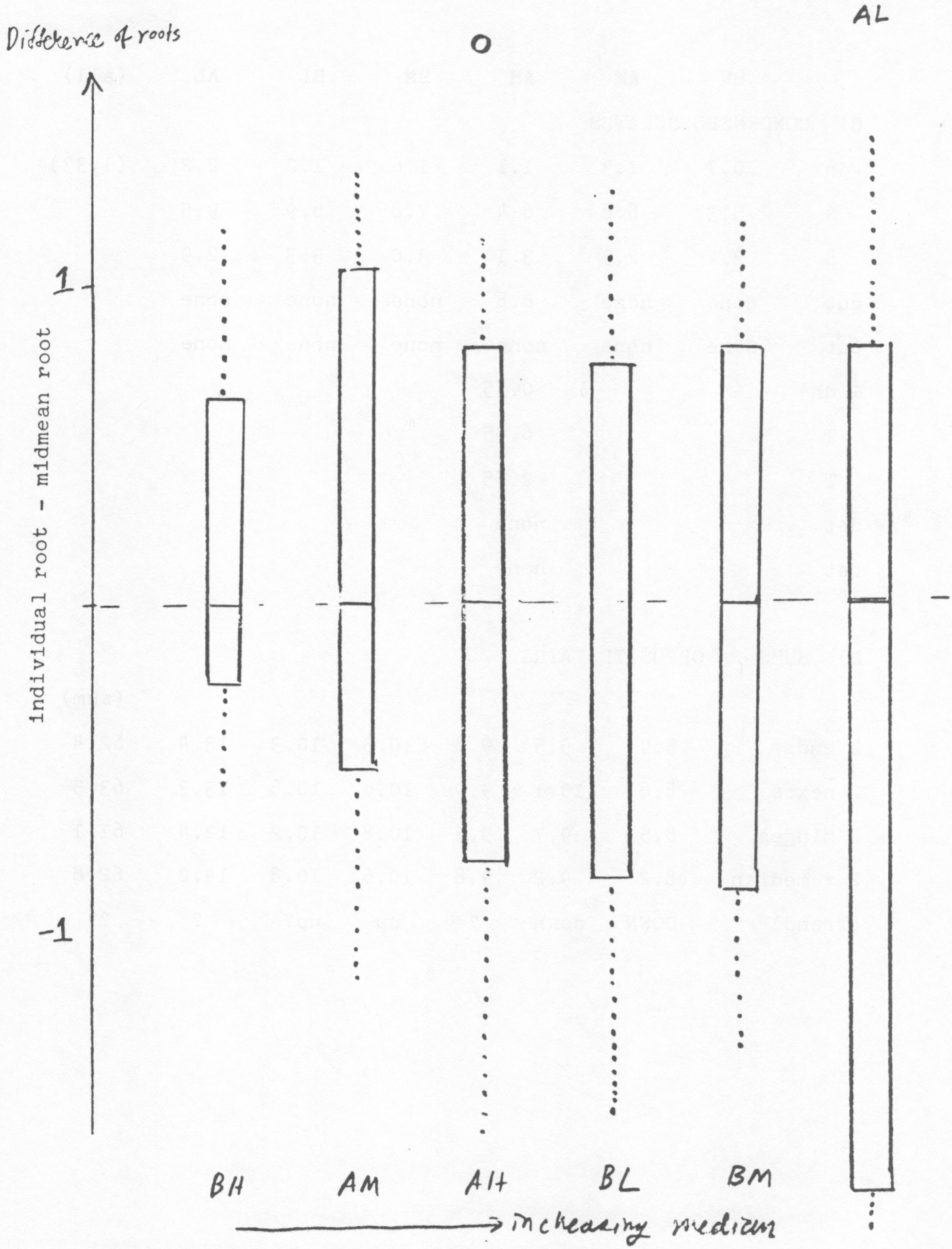
	BH	AM	AH	BM	BL	AL	(all)
C) CONDENSED SCRAWLS							
dh	0.7	1.3	1.1	1.6	1.2	2.2	(1.33)
S	5.3	6.8	6.4	7.8	6.9	9.5	
S	3.2	2.9	3.1	3.0	3.3	2.9	
out	none	none	6.6	none	none	none	
out	none	none	none	none	none	none	
$\frac{1}{2}$ dh			0.55				
T			6.95				
T			2.55				
det			none				
det			none				

D) SUMS of OPPOSITE PAIRS

							(sum)
2 ends	8.9	9.5	9.8	10.5	10.3	13.4	62.4
2 nexts	8.8	10.1	9.9	10.6	10.8	13.3	63.5
2 hinges	8.5	9.7	9.5	10.8	10.2	12.4	61.1
2 x median	8.2	9.2	9.8	10.6	10.8	14.2	62.8
(Trend)	DOWN	down	?	up	up?	?	??

exhibit 9 of chapter 7

Tippett's warp breaks in roots referred to midmean roots



- techniques of looking harder at what expression seems indicated often turn up unsuspected aspects of our data. This happens more often than you would think, and makes real differences in our answers.

- effort in trying any one expression becomes greater and greater the more data is involved.

For all these reasons there is a place for a systematic approach.

(* some artificial examples *)

Let us learn from some artificial examples. Exhibit 10 shows a simple artificial example in which taking logs would provide equal spread, exhibit 11 one in which taking roots would do the same.

The lesson is simple: plot the log of some measure of spread -- say log range or log midspread -- against log median, and look at the tilt of this plot. If

- tilt = 1/2, try roots.
- tilt = 1, try logs.
- (tilt = 3/2, try reciprocal roots).
- tilt = 2, try reciprocals.

If the tilt is indicated very clearly and uniformly, to be intermediate between these values, it may be worthwhile to try some correspondingly intermediate power.

(* warp breaks again *)

Let us return to exhibit 8 and Tippett's warp breaks in square roots, where the ranges, midspreads and medians were:

exhibit 10 of chapter 7

A simple artificial example where taking logs would produce equal spread

A) T EXAMPLE -- five batches wide

End	12	24	36	48	60
Hinge	11	22	33	44	55
Median	10	20	30	40	50
Hinge	9	18	27	36	45
End	8	16	24	32	40
Range	4	8	12	16	20
Midspread	2	4	6	8	10

B) LOG VALUES

log range	.60	.90	1.08	1.20	1.30
log midspread	.30	.60	.78	.90	1.00
log median	1.00	1.30	1.48	1.60	1.70

C) NOTES

A rough plot of either log range or log midspread against log median shows unit slope. (left to reader)

(The fact that range = 2(midspread) is the result of choosing a simple illustration, it will not hold in general.)

exhibit 11 of chapter 7

A simple artificial example where taking roots would produce equal spread

A) T EXAMPLE -- five batches wide

End	144	256	361	484	576
Hinge	121	225	324	441	529
Median	100	196	289	400	484
Hinge	81	169	256	361	441
End	64	144	225	324	400
Range	80	112	136	160	176
Midspread	40	56	68	80	88

B) LOG VALUES

log range	1.90	2.05	2.13	2.20	2.25
log midspread	1.60	1.75	1.83	1.90	1.95
log median	2.00	2.29	2.46	2.60	2.68

C) NOTE

This time, if we plot either log range or log midspread against log median, we get a slope of 1/2. (Again left to reader.)

(Again, range = 2(midspread) is an "accident".)

	BH	AM	AH	BM	BL	AL
Range	1.7	2.5	3.4	2.5	2.9	3.4
Midspread	0.7	1.3	1.1	1.6	1.2	2.2
Median	4.1	4.6	4.9	5.3	5.4	7.1

Exhibits 12 and 13 show the plots of log range and log midspread against log median. In each case the tilt seems to be greater than 1.

A slope of 1 would correspond to going to the logs of the values looked at. Since these were square roots, this would mean going to

$$\log(\text{square root}(\text{count})) = \frac{1}{2} \log \text{count}$$

which is equivalent to going to logs of the counts themselves.

A slope of 2 would correspond to going to the reciprocals of the values looked at. This would mean going to

$$-\frac{1}{\text{square root of count}}$$

which means going to reciprocal roots.

This analysis, then, suggests going at least to logs, and maybe to reciprocal roots, of the original counts.

7E. <u>The meaning of comparison</u>

Much has been said about comparisons. We have used many pictures and many numbers to make comparisons. We have not said what comparison is. It is time we did.

Plot of log range against log median
(tilt somewhat greater than 1)

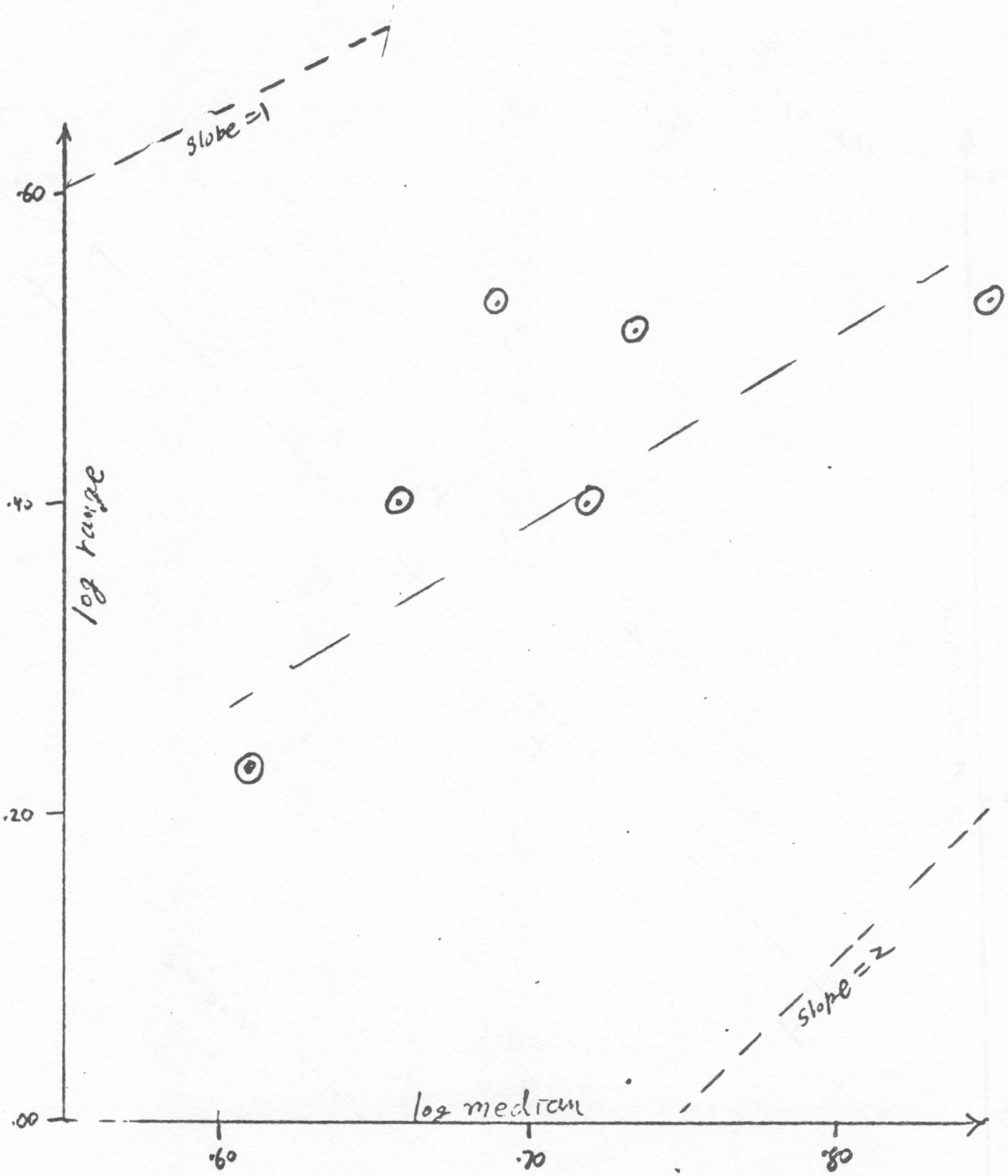
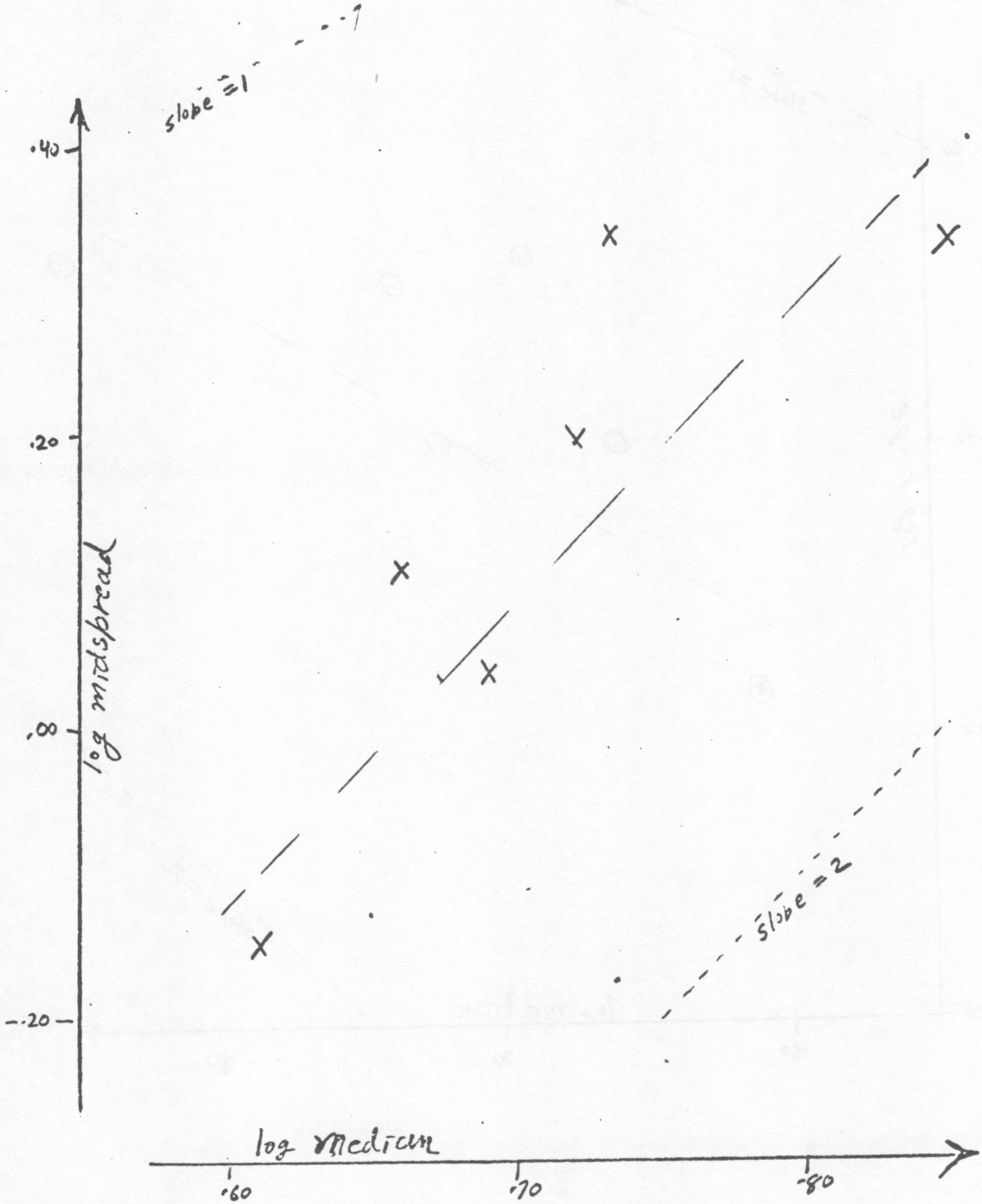


exhibit 13 of chapter 7

Plot of log midspread against log median
(tilt almost 2)



Two kinds of comparisons come up in the simplest of common language. "Bill is a head taller than Jim." "George weighs twice as much as his brother Jack." Each of these statements says what you must do to one person to make him match the other. The first statement bases itself on a plus sign, and says how much we would have to add to Jim's height to match Bill's. The second bases itself on a times sign, and says what we would have to multiply Jack's weight by to match George's.

There are many ways in which addition is simpler than multiplication: doing hand arithmetic and sliding things bodily along graphs are only two. Logarithms were invented centuries ago to reduce multiplication to addition, thus making hand arithmetic much easier. We can, do, and should use them to avoid comparisons by multiplication. When multiplication would seem needed, taking logs will let us deal with comparison by addition.

When we think of comparison, then, we want to think of what needs to be added or subtracted to the data -- as carefully expressed for analysis -- in order to make one thing match another. If the data is originally, or commonly, expressed in another way, we will try to translate back to that expression too.

Thus, if boys' weights are really better thought of as "twice" or "three-quarters" of one another, we will do our

analysis in log weight, finding, perhaps, that we have to add 0.30 to Jack's to match George's. Having found this, we would also convert back to raw weight, and say that Jack's had to be doubled in order to match George's. Both forms of answer help.

7F. Rough adjustments and rough residuals

If we want to put several batches under the same microscope at the same time, we need to have them centered at about the same level. If they are not naturally similarly centered, we will have to adjust them so that they will be.

Thanks to the flexibility of the human eye it is not important -- if we are to look at a picture -- that the centering be exactly the same. Rough adjustments will serve us essentially as well as precise ones. We can form rough residuals and be happy with their behavior.

(* plankton again *)

Let us then turn again to the Winsor and Clarke data. Exhibit 7 showed, in raw terms, that Kind III is roughly three times as common as Kind IV, which is roughly 15 times as common as Kind I. One way to look harder at the data is to form a rough match by making rough adjustments. Exhibit 14 gives pushed-together numbers, while exhibit 15 shows an expanded view of both individual values and schematic plots on a log scale. It is clear that this kind of adjustment, which we want to think of as adjustment by addition or

exhibit 14 of chapter 7

The Winsor and Clarke numbers pushed together
for closer comparison
(see exhibit 7)

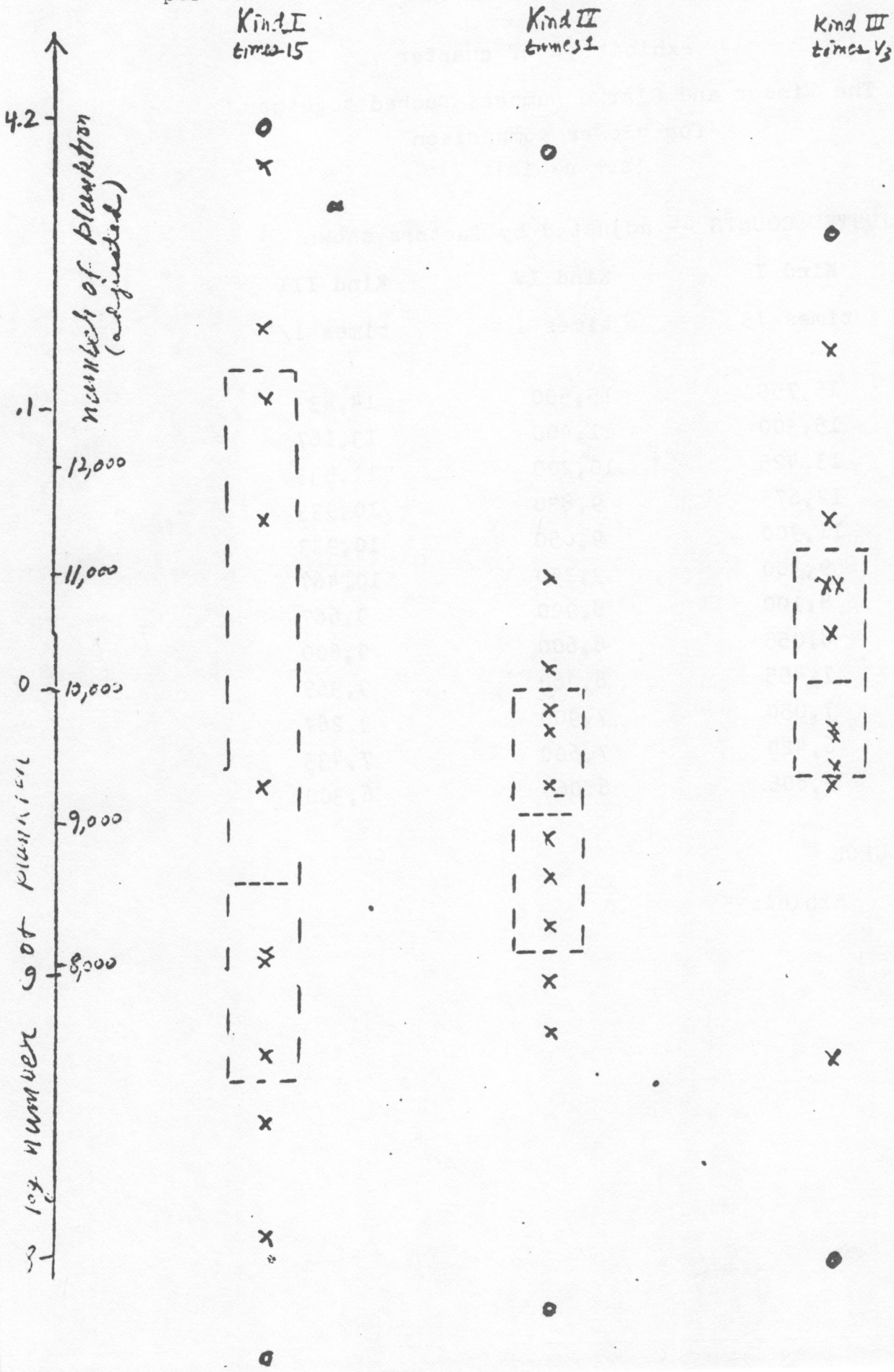
A) ADJUSTED COUNTS -- adjusted by factors shown

Kind I	Kind IV	Kind III
times 15	times 1	times 1/3
15,750	15,500	14,433
15,300	11,000	13,167
13,425	10,200	11,533
12,675	9,830	10,933
11,900	9,650	10,933
9,300	9,250	10,467
8,100	8,900	9,667
8,055	8,600	9,600
7,455	8,260	9,367
7,050	7,900	9,267
6,420	7,600	7,433
5,805	6,060	6,300

B) SOURCE

Exhibit 5

Winsor and Clarke's estimated numbers of plankton per haul of net made comparable by adjusting factors



subtraction on a log scale, has left the three sets of numbers in a very good state to compare in detail. As examples, we can see that:

- even after adjustment, Kind III seems about 10% higher than the others.
- there may be a gap in the middle of the Kind I values.

Had we tried to compare the raw numbers similarly, we would have had to add, say, 8500 to the numbers for Kind I and subtract, say, 20,000 from the numbers for Kind III, obtaining the results shown in exhibit 16. Only the schematic summaries are shown, because we could not plot the 12 adjusted values for Kind I and have them individually visible. Detailed comparison is almost useless.

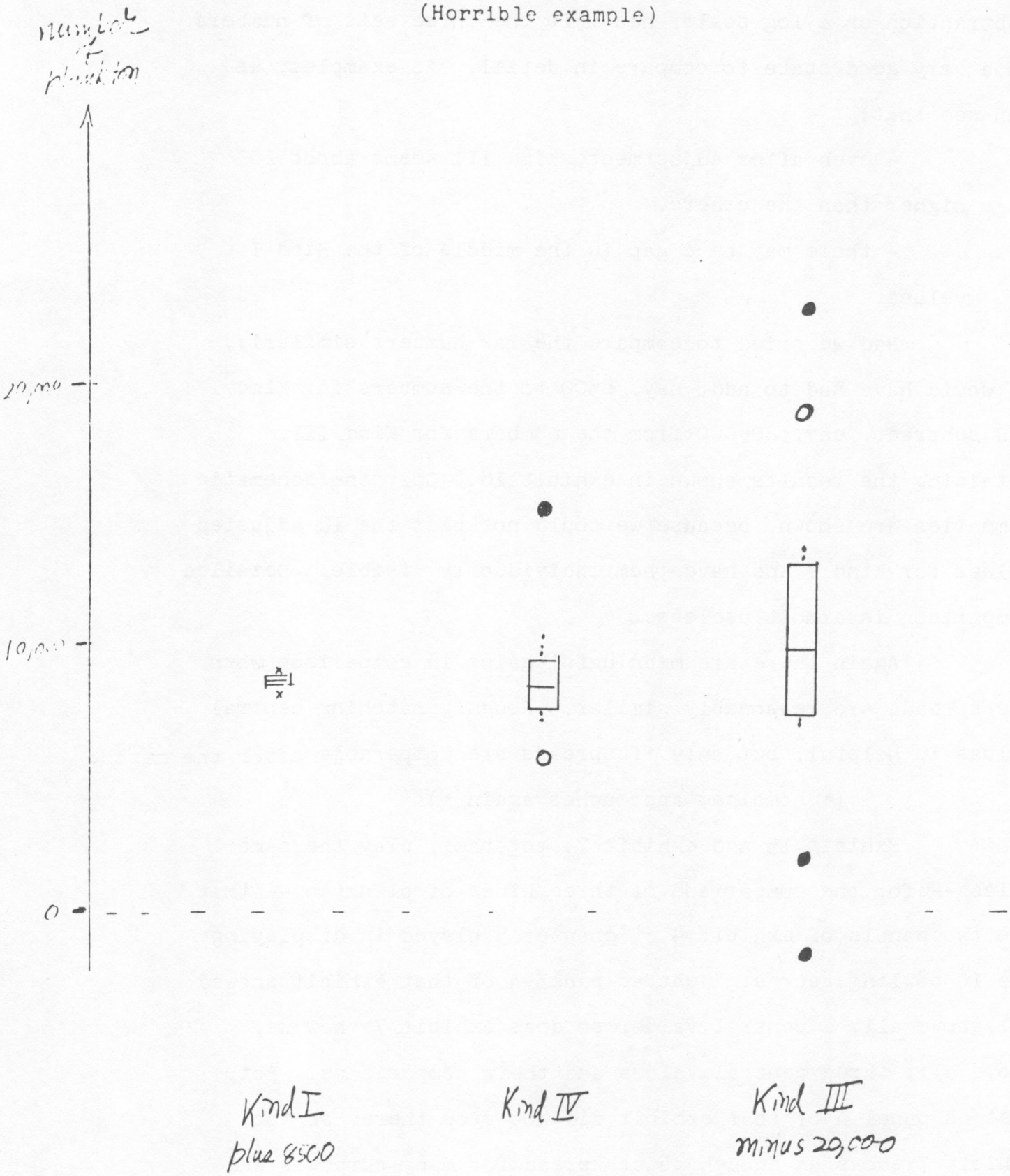
Again there are meaningful gains in comparison when the spreads are reasonably similar. Roughly matching central values is helpful, but only if spreads are comparable after the match.

(* combined approaches again *)

Exhibit 15 and exhibit 7, together, play the same roles -- for the comparison of three kinds of plankton -- that the two panels of exhibit 4 of chapter 5 played in displaying the 10 bowling scores. Just as panel A of that exhibit showed us, above all, a central value, so does exhibit 7 show us, above all, three central values and their comparisons. But, just as panel A of that exhibit did not stop there, so too exhibit 7 shows us enough about spread for many purposes and a little about pattern.

Winsor and Clarke's estimated numbers of plankton per haul of net made comparable by additive adjustments

(Horrible example)



Just as panel B of that exhibit showed us both the individual values and a muted form of their middle-and-ends summary, so that we could concentrate on pattern, either in detail or in summary, so too exhibit 15 does this for three batches -- batches whose raw values are so different that we could not see these patterns without bringing the batches together by rough adjustment.

If we are to look, it is not important whether matching is precise, we do not need exact central values and exact residuals. The eye can take up the slack. But it can be essential to use rough adjustments to allow us to blow up the picture enough to see all relevant detail -- to gain what rough residuals have to offer.

7G. <u>The pursuit of symmetry takes much data</u>

Sometimes we can gain by re-expressing our data to make their pattern more symmetrical. Suceessfully gaining symmetry can reduce the number of apparent strays, sometimes markedly.

(* back to volcanoes *)

Let us go all the way back to chapter 3, where exhibits 7, 8 and 9 set out the heights of 218 volcanoes raw, in logs, and in roots. Exhibit 6 of chapter 5 gave the corresponding schematic plots. Exhibit 17 gives the condensed summaries to remind us of the general situation. Clearly neither raw nor log heights are symmetrical, although root heights might be. (The cube root might do even better.)

exhibit 17 of chapter 7

First-stage skipping analyses of three expressions of the same data

(see exhibits 7 to 9 of chapter 3)
(plots as exhibit 6 of chapter 5)

A) RAW HEIGHTS -- for 218 volcanoes, in thousands of feet

```
218|U
  H|  3.7
  M|  6.6
  H|  9.6
tri|  6.6

 dh|  5.9 up 3.0
S/T|- 2.2 / -5.2
S/T| 15.5 / 18.5
out|  (0,0)
out|(13,8)
```

B) LOG HEIGHTS -- for the same

```
218|U
  H|1.58
  M|1.82
  H|1.98
tri|1.80

 dh|0.40 up 0.20
S/T|1.18 / 0.98
S/T|2.38 / 2.58
out|(11,8)
out| (0,0)
```

C) ROOT HEIGHTS -- for the same

```
218|U
  H|62
  M|80
  H|98
tri|80

 dh| 36 up 18
S/T| 26 / 8
S/T|134 / 152
out|(5,0)
out (2,0)
```

With enough data, a batch can tell how it should be expressed to reach symmetry. In this example, 218 volcanoes seem almost enough to settle matters. Those who go on to read chapter 7x will find that 2,306,363 incomes seem to be enough.

If we are to look for hints about symmetry in more marginal cases, we need a simpler and more systematic approach. Let us try to use as many of the simple summary values -- median, hinges, depth-2s (i.e. values whose depth is 2) extremes -- as we have available. Symmetry would surely be indicated if all these values were the same:

- median
- mean of hinges
- mean of depth-2s
- mean of extremes

Most of these values are means of 2. It is less work to multiply by 2 once than to divide by 2 more than once. So we may as well look at the following set of values, each exactly twice the corresponding value in the last list:

- 2 times median
- sum of hinges

- sum of depth-2s
- sum of extremes

Exhibit 18 applies this technique to the 216 volcanoes in their 3 previous expressions, and one more. The analysis of raw heights, logs of heights and square-roots of heights merely strengthens the visual impression from exhibits 7, 8, and 9 of chapter 3:

- square-roots do far better than the other two.
- square-roots do very well indeed

The fourth column, (*), in exhibit 18 is for cube-roots. Clearly square roots are much better.

(* precipitation again *)

We can apply the same technique to several batches rather than to a single batch. We can illustrate at the same time both

- how to look a several batches
- how only a few small batches with small greatest/least ratios can tell us almost nothing.

Exhibit 19 returns to the New York City precipitations examined in exhibits 1 to 4, and does just this. We look at each batch separately, as before, and add up corresponding terms across batches. For the raw values we get a rather confusing result. If we were dealing with the ordinary type of asymmetry, we would expect the sum of hinges close to twice the median, and the sum of extremes further away. Actually,

exhibit 18 of chapter 7

Looking at 218 volcanoes for symmetry of pattern

A) SUMMARY VALUES

	Raw	Logs	Roots	(*)
extreme	0.2	0.30	14	5.8
depth-2	0.5	0.70	22	7.9
hinge	3.8	1.58	62	15.5
median	6.6	1.81	80	18.8
hinge	9.6	1.98	98	21.2
depth-2	19.7	2.29	136	27.0
extreme	19.9	2.30	140	27.1

B) INDICATORS SYMMETRY

2 times median	13.2	3.62	160	37.6
sum of hinges	13.4	3.56	160	36.7
sum of depth-2s	20.2	2.99	158	34.9
sum of extremes	20.1	2.60	154	32.9
(trend?)	(UP)	(DOWN)	(down?)	(down)

C) REMARKS

(*) gives cube roots, as suggested by a less adequate analysis.

exhibit 19 of chapter 7

What expression makes the pattern of New York City annual rainfall symmetrical?

A) TRIAL of RAW VALUES

	189-	190-	191-	192-	193-	194-	195-	(sum)
2 times median	85	91	78	75	75	91	88	(583)
sum of hinges	91	89	77	80	80	90	87	(594)
sum of extremes	89	91	78	83	83	88	85	(597)

B) TRIAL of LOG VALUES

2 times median	3.25	3.31	3.18	3.15	3.15	3.31	3.28	(22.63)
sum of hinges	3.31	3.29	3.17	3.20	3.19	3.30	3.27	(22.73)
sum of extremes	3.28	3.31	3.17	3.22	3.22	3.28	3.25	(22.73)

C) COMPARISON OF DEVIATIONS

(panel A minus 80)

5	11	-2	-5	-5	11	8
11	9	-3	0	0	10	7
9	11	-2	3	3	8	5

(100×panel B minus 320)

5	11	-2	-5	-5	11	8
11	9	-3	0	-1	10	7
8	11	-3	2	2	8	5

D) SAMPLE CALCULATION

For 1890's = 52, 41, 39, 53, 44, 36.
Ordered = 36, 39, 41, 44, 52, 53.
Sums = 41 + 44 = 85; 39 + 52 = 81; 36 + 53 = 89.
Logs (of ordered) = 1.56, 1.59, 1.61, 1.64, 1.72, 1.72.
Sums = 1.61 + 1.64 = 3.25; 1.59 + 1.72 = 3.31; 1.56 + 1.72 = 3.3

the sums of hinges sum to 594 -- and the sum of extremes to 597 -- while twice the median is further away at 583. As panel C stresses, the effect of taking logs is very nearly linear.

We can only conclude that it would take a lot more data of this sort to give any clear indication as to what expression promotes better symmetry. Fortunately, we can be even more sure that any change in symmetry due to a change of expression is very unlikely to be of any importance to us so far as this particular sort of precipitation data goes. Other lengths of accumulation or other locations may lead to clear choices of expression for symmetry and to some clear usefulness for such choices, but no trace of either is to be seen here.

(* a comment *)

Why should this section -- which seems quite explicitly to do with individual batches -- appear in a chapter on comparisons? There are reasons, strange though this may seem.

First, we more usually change expression to promote comparison. Getting clearer pictures or simpler descriptions by removing unnecessary differences in spread is a larger gain. We depend on this as the first consideration in dealing with batches of numbers.

It is only when it fails to urge us to choose one form of expression rather than another that we turn to symmetry as last resort. This failure can come because of some or all of:

- the batches are all at or near the same level
- the batches are small enough for the variation in individual spread natural in small batches to conceal the effects of level.

Whether we like it or not, comparison determines whether or not we look at a batch or a set of batches to see what light symmetry can turn on choice of expression.

Secondly, inquiry into symmetry is a kind of comparison -- a comparison of the lower half of the batch with the upper half, as seen from the middle.

Third, and most important of all, while it would be quite possible for symmetry to prefer one choice of expression for the same sort of data that constant spread prefers another, this just

does not seem to happen

Messages about choice of expression, whether the more important messages about avoiding unnecessary shift of spread or the less important messages about gaining symmetry, almost always agree, given enough data.

This is a clear empirical fact. (We can also even give a plausible explanation, but not within the scope of this account.)

All in all, this is the right chapter to discuss the message from symmetry, because of relationship, nature, and detailed agreement.

*****

7H. How far have we come?

We opened this chapter with a realistic example of how comparison of single numbers sometimes tells the whole story, and with other examples of how it sometimes misses it completely. We cannot fail to use single-number comparisons, but we can be much happier about doing so if we have looked at -- and compared -- broader summaries as well, whether or not we report that look.

Comparisons are indeed everywhere. A few illustrations were given.

The simplest and most frequent opportunity to make comparisons more useful is to avoid unnecessary confusion by eliminating unnecessary change of spread with level, by choosing the way the data are expressed. Doing this can greatly increase clarity and strengthen comparison.

We can find a good expression by trial and error. With more data, though, it helps to use a systematic approach, plotting the log of a simple measure of spread against the log of a simple measure of level, and using the apparent tilt as a guide both as to direction and as to amount of change.

Comparison, in common language, is a matter of difference or a matter of ratio. Logs were invented to bring ratios to differences.

If we want more than one batch under a single microscope, if we want to see their details together, rough

adjustments can serve us very well. For visual comparison, high residuals again are almost as good as precise ones.

If there is no other reason to choose expression, we can seek symmetry. Large enough batches with large enough ratios like

$$\frac{\text{greatest}}{\text{least}} \quad \text{or} \quad \frac{\text{high hinge}}{\text{low hinge}}$$

will tell us which choice favors symmetry (if such a choice there be). This type of inquiry is more closely related to comparison than we would think.

*****

7K. Effects of re-expression (background)

What happens when we change expression? We are often concerned with quantities that are particular observed values picked out by their rank among the ordered observed values. These include:

- end values
- next-to-end values
- hinges (except those for batches of 4,8,12,...)
- medians (of batches of an odd number of values).

These values change in just the same way as a single observed value -- at least for the kinds of change of expression we have considered.

Next to them fall:

- hinges (in batches of 4,8,12,...)
- medians (of batches with an even number of observations)

which are means of two observations, to be sure, but are means of two observations adjacent in value within the batch, and to that extent as close as we can get them. Next come means of rather more widely spread values (usually middle means):

- trimeans
- midmeans

These are all means of closely packed values. They do not behave quite so simply, but they come close.

Finally, we have means of quite widely separated values and differences of at least fairly widely separated ones, including:

- the mean
- the midspread
- the range

The behavior of these quantities when the basic values are re-expressed is nowhere nearly as simple.

Some specific examples may help to fix these ideas. Exhibit **20** shows two simple examples, one of change to (square) roots, the other of change to logs. When we look at the % differences -- in % where appropriate -- between:

- a quantity calculated from raw values, and
- the results of re-expressing in raw form the same quantity calculated from re-expressed values,

both examples show a steady progression from no difference to large differences, namely:

- unchanged quantities
- median (when found from two observations)
- hinges (when found from two observations)
- midmean
- trimean
- mean
- range and midspread

This progression follows in broad outline the order suggested by the discussion above; carrying the discussion further would predict the finer detail.

In general, for data with largest/smallest ratios that are substantial but not tremendous, we can often go as far as the midmean without serious concern, and somewhat less

exhibit 20 of chapter 7

Some simple examples of the effect of re-expression

A) T SUPPOSED DATA -- 18 to 1 for largest to smallest, re-expressed as roots

Raw	Root	Restored (Last Col.) (Squared)	Change (Amount)	Percen
160	13	169		
127	11	(121)		
100	10	(100)	-	
64	8	(64)	-	
49	7	(49)	-	
36	6	(36)	-	
25	5	(25)	-	
16	4	(16)	-	
9	3	(9)	-	

B) DERIVED VALUES

end	169	13	(169)	-	
next	121	11	(121)	-	
hinge	100	10	(100)	-	
median	49	7	(49)	-	
hinge	25	5	(25)	-	
next	16	4	(16)	-	
end	9	3	(9)	-	
midmean	49.7	7.00	(49.0)	-0.7	-1%
trimean	55.8	7.25	(52.6)	-3.2	-6%
mean	64.3	7.44	(55.3)	-9	-14%
midspread	75	5	(25)	-50	-67%
range	160	10	(100)	-60	-38%

C) SECOND EXAMPLE -- 62 to 1 for largest to smallest, re-expressed as logs

Raw	Log	Restored (Last Col.) (un-logged)	Change
2512	3.40	(2512)	-
1000	3.00	(1000)	-
631	2.80	(631)	-
501	2.70	(501)	-
398	2.60	(398)	-
251	2.40	(251)	-
126	2.10	(126)	-
40	1.60	(40)	-

(Cont'd)

exhibit 20 (Cont'd)

D) DERIVED VALUES

end	2512	3.40	(2512)	-	
next	1000	3.00	(1000)	-	
next	126	2.10	(126)	-	
end	40	1.60	(40)	-	
hinge	816	2.96	(794)*	(-22)**	
median	450	2.65	(447)*	(-3)**	
hinge	188	2.25	(178)*	(-10)**	
midmean	445	2.625	(422)	-23	-5%
trimean	476	2.612	(410)	-66	-14%
mean	682	2.575	(376)	-206	-30%
midspread	628	.65	(4.5)	-624	-99%
range	2472	1.80	(63)	-2409	-97%

* Differ from first column, since logs in second column are means of logs of adjacent values.

** Not zero because of * .

often as far as the trimean. The mean is likely to be seriously altered.

If the largest/smallest is close to 1, everything will be well behaved. If it is tremendous, nothing that is not unchanged will behave simply.

7L. A psychological example (optional)

In 1950, Bruner, Postman, and Mosteller reported on an experiment to study and clarify a psychological phenomenon, the reversal of perspective. Certain pictures can be seen as in either of two perspectives; most subjects see first one, then the other, so that it is natural to measure how rapidly the perspective reverses. In the experiment this was done separately for each minute of a 10-minute trial, 19 subjects being studied under each of three sets of instructions:

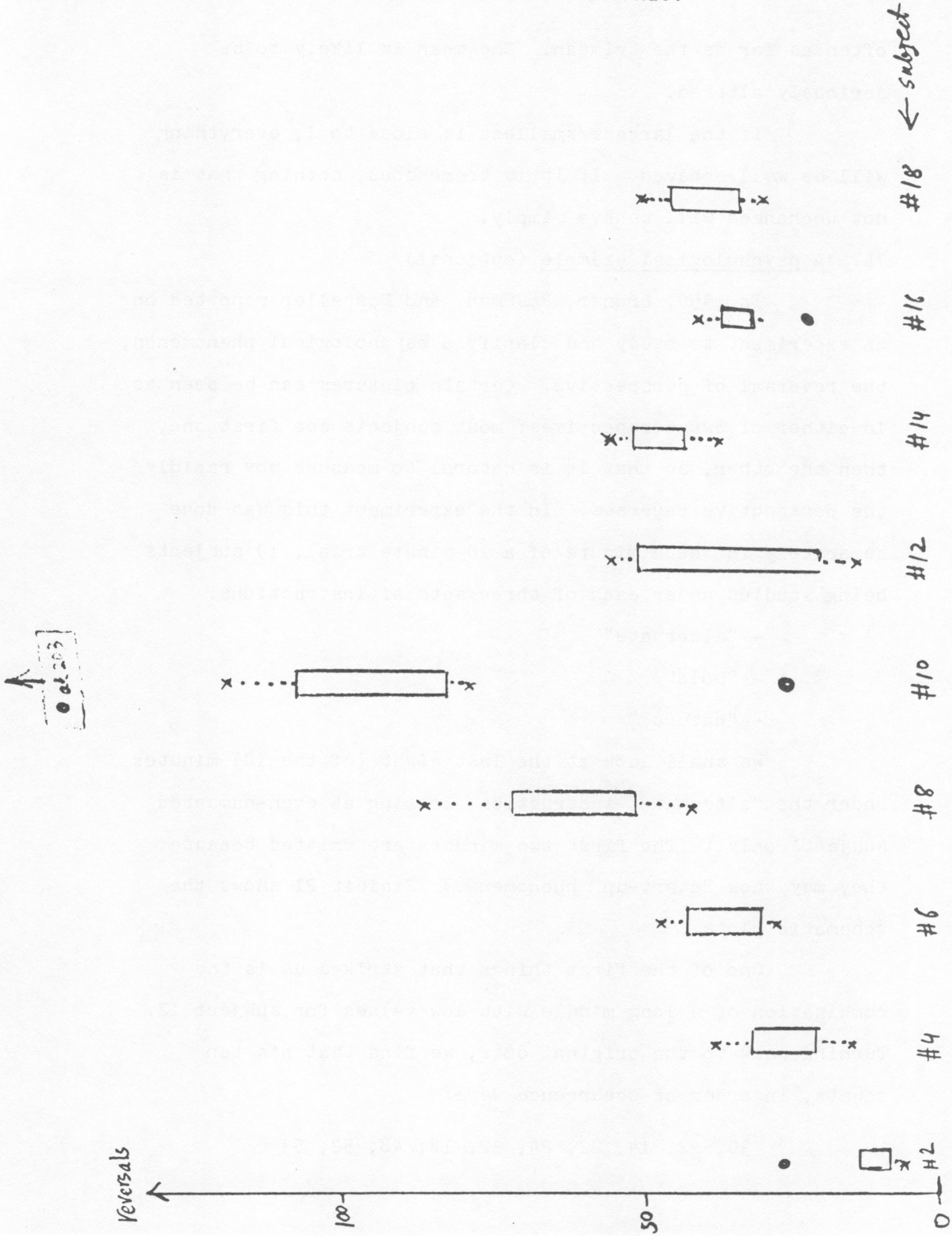
- "alternate"
- "hold"
- "natural"

We shall look at the last eight (of the 10) minutes under the "alternate" instruction, looking at even-numbered subjects only. (The first two minutes are omitted because they may show "start-up" phenomena.) Exhibit **21** shows the schematic plots.

One of the first things that strikes us is the combination of a long middle with low values for subject 12. Turning back to the original data, we find that his ten counts, in order of occurrence were:

30, 22, 14, 22, 24, 32, 18, 48, 52, 53

Schematic Plots for 9 Subjects (Even Numbers) When Instructed to "Alternate"



It would appear that after seven minutes this subject suddenly learned how to make his perspective reverse rapidly. We shall have to allow for this as we proceed.

Since the typical number of alternations varies widely, we naturally wish to see whether other characteristics of the counts vary with it. Exhibit 22 shows the same displays for all even-numbered subjects except #12, with the subjects ordered according to midmean. This plot strongly suggests that:

- larger spreads appear with larger midmeans.
- the lowest and highest values for subject 10 are anomalous, even when we allow for the large midmean.

To look more closely at the spreads of the various individuals, we turn to exhibit 23, which plots midspread against midmean for each batch of eight counts. The suggestion of a relationship appears strengthened.

The natural ways to re-express counts are, we believe, roots or logs. Exhibit 24 and exhibit 25 show the midspreads for these expressions of the data. In these plots, in order to simplify and clarify comparing one plot with another, we have used the midmeans of the raw values as the horizontal scale. We see that:

- the spreads of the square roots still seem to trend upward.
- the spreads of the logs seem to trend downward if they trend at all.

exhibit 22 of chapter 7

Schematic Summaries Ordered by Midmean
for 8 of the Same Subjects

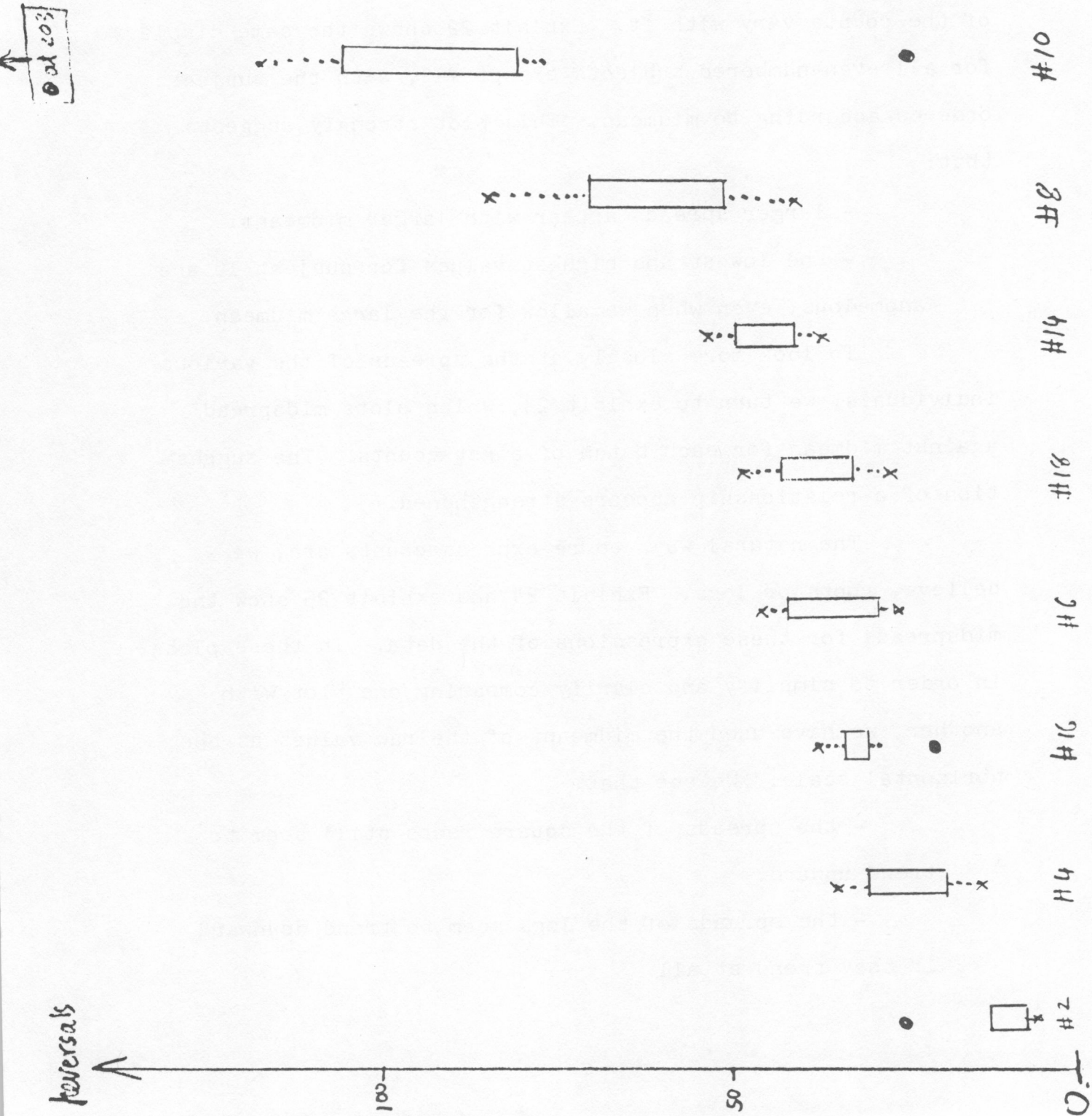


exhibit 23 of chapter 7

Midspread plotted against midmean
(for all 9 subjects, #12 plotted as "12")

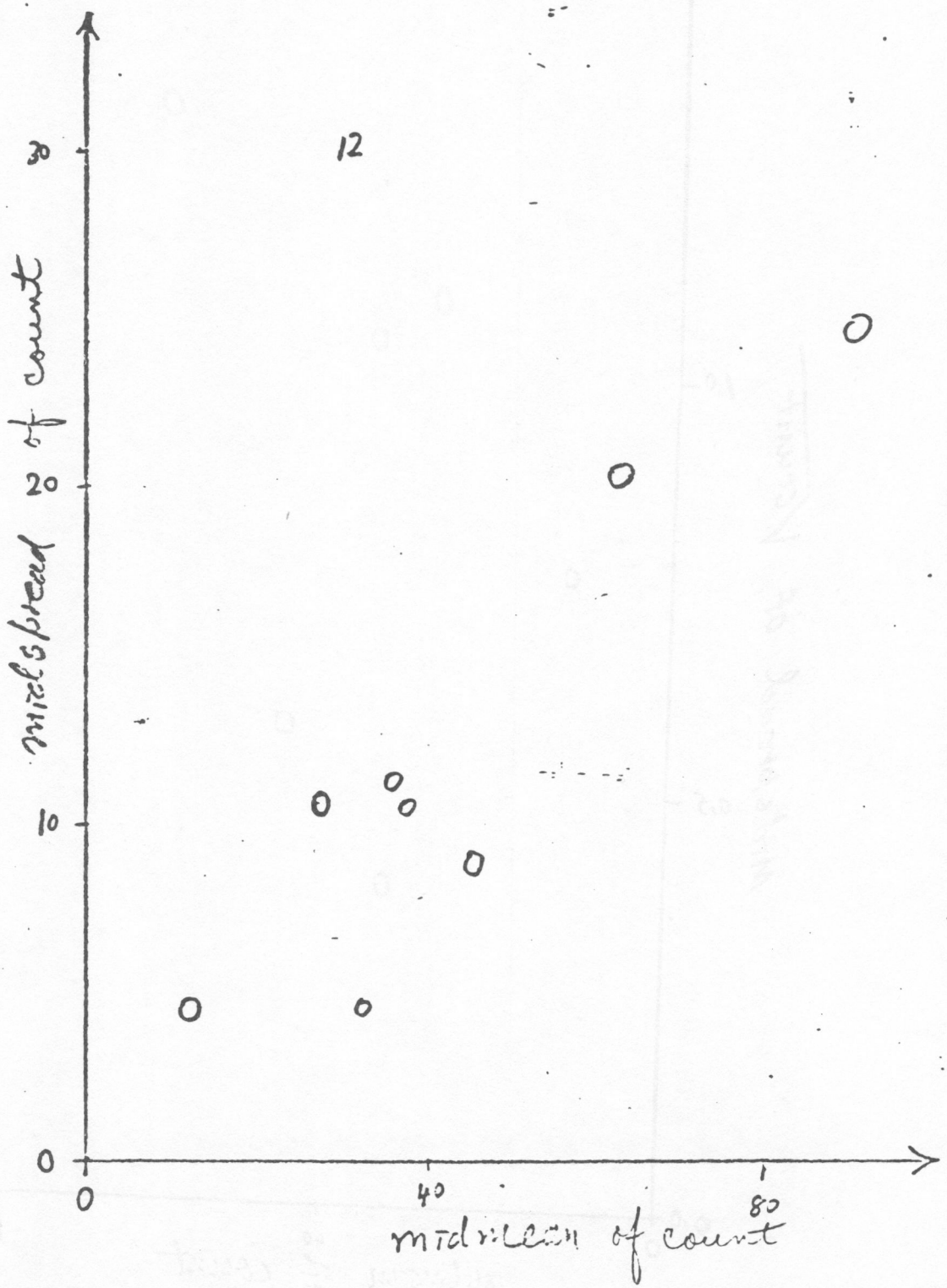


exhibit 24 of chapter 7

Midspread of log count against midmean
(midmean of count)
for the same 8 subjects
(#12 omitted)

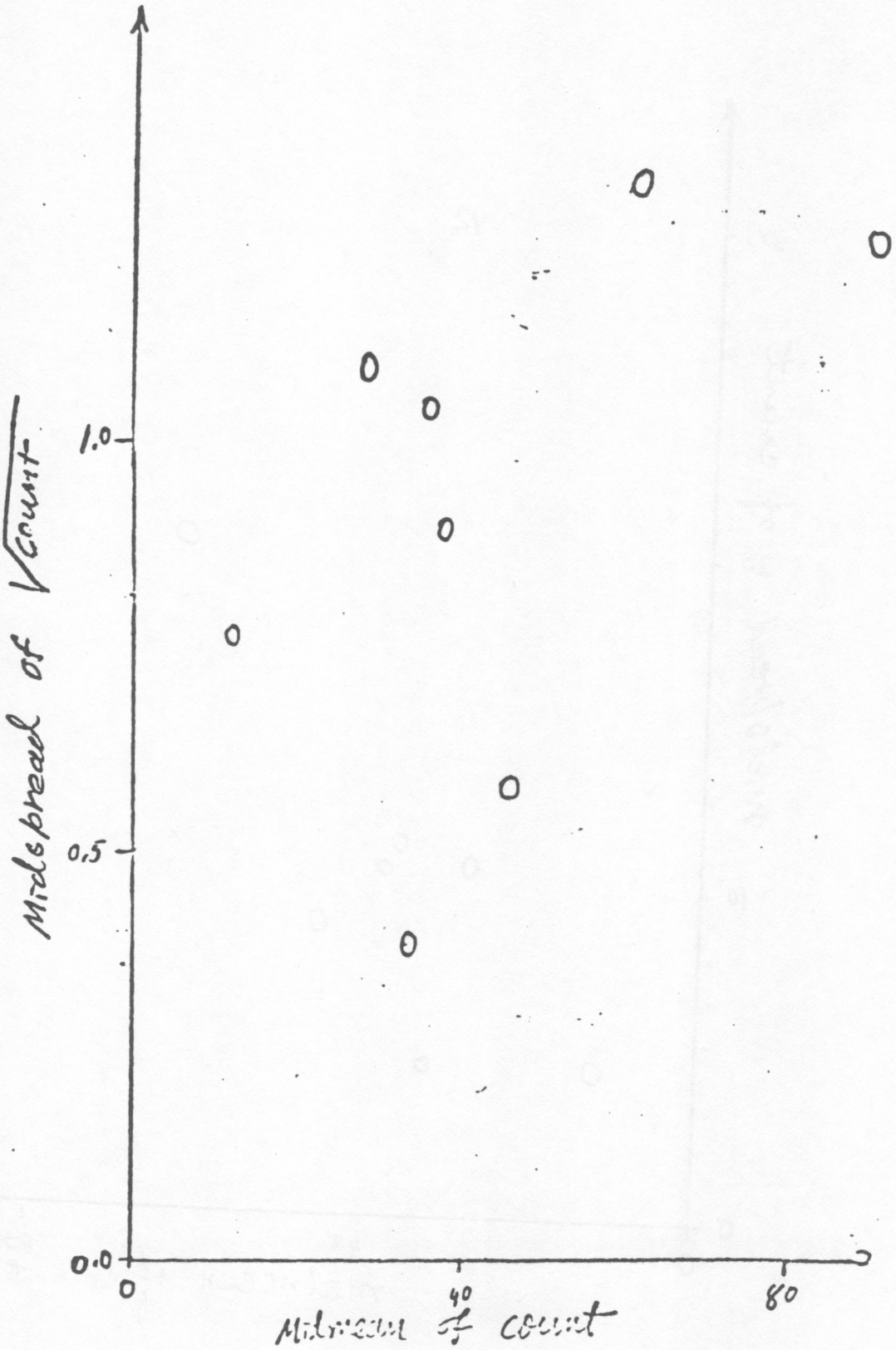
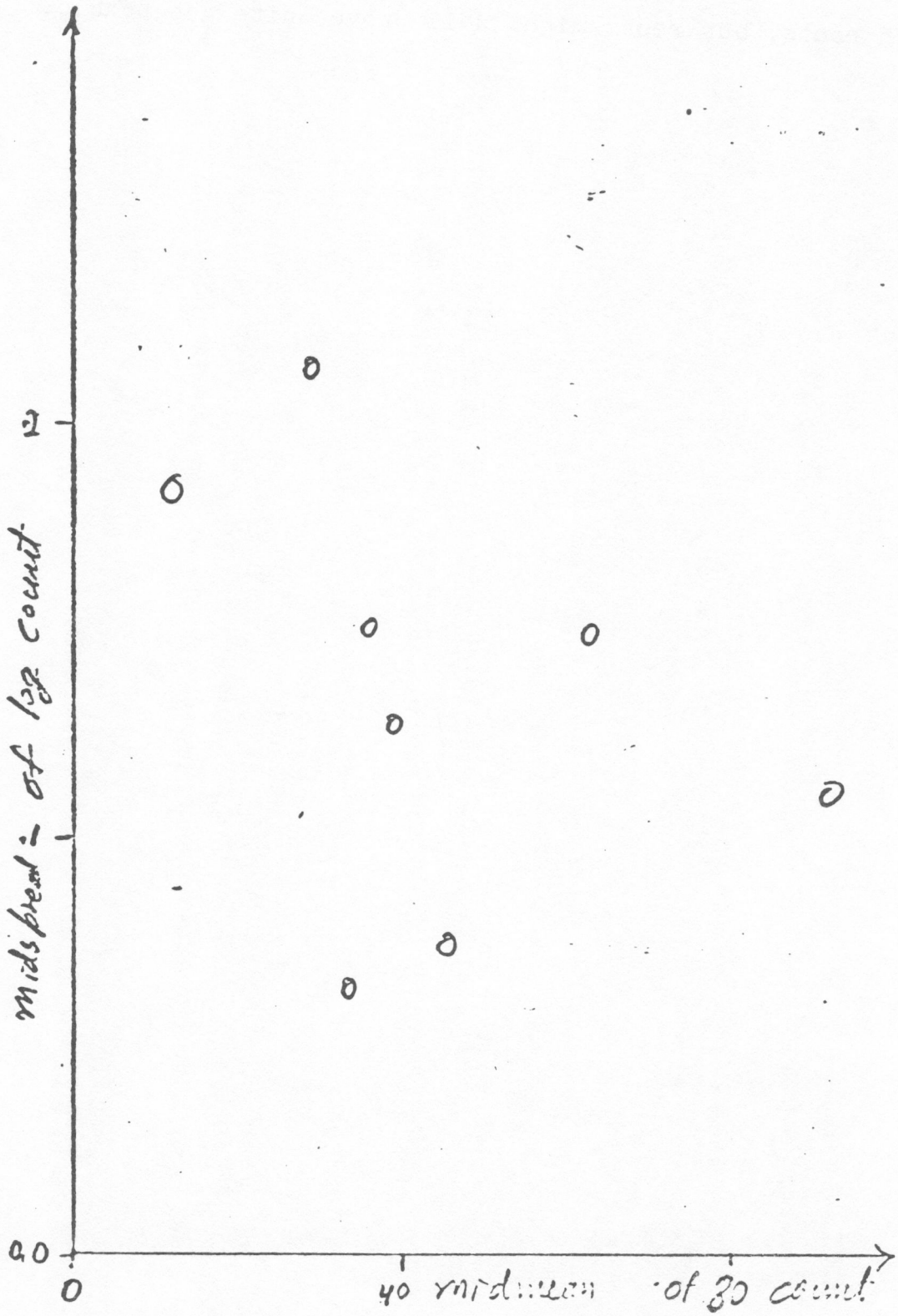


exhibit 25 of chapter 7

Midspread of log count against midmean for the 8 subjects (omitting #12). (Midspread of log counts)



Some change in expression is indicated, though the amount of data we have been exploring is not enough to give a clear indication. Logs look a little preferred as compared with roots, but roots might well prove quite acceptable.

7P. Problems for chapter 7

1. Exhibit 1 gives New York City precipitation for seven batches of six consecutive years. Calculate ranges, next ranges, and midspreads, and make schematic plots. Comment.

2. Are there any outside values in any of these seven batches?

3. Problems (1) and (2) of chapter 4 give New York City snowfalls for two batches, each of 20 consecutive years. Array as in exhibit 3, calculate the measures of spread shown in that table, and make schematic plots. Comment.

4. What light do these data throw on how they should be expressed?

5. The snowfalls in Buffalo, New York and Cairo, Illinois for the 20 consecutive winters from 1918-19 to 1937-38 are given in exhibit 26. Make schematic plots for the three places, including New York City, for this time period.

6. What light do these three batches throw on how they should be expressed?

7. Exhibit 14 of chapter 4 gives, for two elections, % of major party vote for Republication presidential candidates in eight southwestern states. Make schematic plots for:

- the 1960 data,
- the 1964 data,
- the change (swing) from 1960 to 1964.

Comment.

exhibit 26 of chapter 7

Seasonal Snowfall in Buffalo, New York

and Cairo, Illinois from 1918-19

to 1937-38 (inches)

	Buffalo	Cairo
1918-19	25.0	1.8
1919-20	69.4	4.5
1920-21	53.5	13.9
1921-22	39.8	4.0
1922-23	63.6	1.2
1923-24	46.7	6.8
1924-25	72.9	7.2
1925-26	79.6	11.5
1926-27	83.6	6.2
1927-28	80.7	0.4
1928-29	60.3	11.5
1929-30	79.0	12.4
1930-31	64.8	11.3
1931-32	49.6	2.9
1932-33	54.7	7.4
1933-34	71.8	2.7
1934-35	49.1	1.6
1935-36	103.9	14.1
1936-37	51.6	5.4
1937-38	81.6	3.0

Sources: Report of the Chief of the Weather Bureau 1918-1919 to 1934-1935, and U. S. Meteorological Yearbook 1935 to 1938.

8. Problem (6) of chapter 4 gives % carotene in the flour for 99 varieties of wheat. Treating these 99 values as a single batch, make schematic plots. Comment. Are there any outside values? Do you think some other form of expression would be better?

9. Beall (1942) counted the number of Phlegethontius quiquemaculata (an insect) per plot after various insect control treatments. Exhibit 27 gives a simplified form of his data. What choice of expression is indicated if we take the counts for each treatment as a batch?

10. In 1948, an anonymous correspondent gave the weights of ducks after various protein supplements in exhibit 28. Make an effective graphical presentation and comparison.

11. In 1949, Geiger and Dunn gave the weight gains of rats in exhibit 29. Make an effective graphical presentation and comparison.

12. In 1946, Brade gave the electrometric iodine numbers in exhibit 30. Make an effective graphical presentation and comparison.

13. In 1948, Weaver and Riley gave the data in exhibit 31 concerning water permeability of membranes. Make an effective graphical presentation and comparison.

14. In 1945, Isbin and Kobe gave the results in exhibit 32. Make the most effective graphical presentation and comparison you can.

15. In 1952, Hersh and Montgomery gave the electrical resistivities of fibers in exhibit 33. Make the best graphical comparison and presentation you can.

exhibit 27 of chapter 7

Number of insects per plot after various insect-control treatments

(data simplified)

Treatment					
1	2	3	4	5	6
10	11	0	3	3	11
7	17	1	5	5	9
20	21	7	12	3	15
14	11	2	6	5	22
14	16	3	4	3	15
12	14	1	3	6	16
10	17	2	5	1	13
23	17	1	5	1	10
17	19	3	5	3	26
20	21	0	5	2	26
14	7	1	2	6	24
13	13	4	4	4	13

Source: Geoffrey Beall 1942. The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. Biometrika 32:243-262

exhibit 28 of chapter 7

Weights of chicks after various protein supplements

A) The WEIGHTS in GRAMS at age SIX WEEKS

Protein supplement fed					
Horse -bean	Linseed oil meal	Soybean oil meal	Sunflower seed oil meal	Meat meal	Casein
179	309	243	423	325	368
160	229	230	340	257	390
136	181	248	392	303	379
227	141	327	339	315	260
217	260	329	341	380	404
168	203	250	226	153	318
108	148	193	320	263	352
124	169	271	295	242	359
143	213	316	334	206	216
140	257	267	322	344	222
	244	199	297	258	283
	271	171	318		332
		158			
		248			

B) The SOURCE

(Anonymous) 1948. Query 60. Biometrics 4: 213-214.

exhibit 29 of chapter 7

Weight gain of rats with different dietary supplement

A) The GAINS in GRAMS over a 16-DAY PERIOD

Dietary supplement				
Gluten	Gluten and lysine	Gluton and α-aminoadipic acid	Casein	Casein and α-aminoadipic acid
6	18	1	34	31
4	30	-4	30	43
-2	33	6	39	39
2	34	-3		35
3	31	-1		32
1	28	0		40

B) The SOURCE

E. Geiger and H. J. Dunn 1949. On the failure to utilize α-aminoadipic acid as a precursor of lysine by rats and certain bacteria. Journal of Biological Chemistry 178: 877-879.

exhibit 30 of chapter 7

Electrometric iodine numbers

A) The DATA

Substance*	Electrometric iodine numbers
Cyclohexene	306,296,302,299,312,312,306,314,
2-octene	212,209,213,212,
1-octadecene	98.2,99.5,97.8,99.2,
styrene	224,209,217,212,208,218,220,
2-methyl-p-methyl styrene	204,176,187,187,187,179,
pinene	268,246,234,238,217,219,
isoeagenot methyl ether	135,133,133,133,134,

* 18 substances in the original

B) The SOURCE

B. Brade 1946. Determination of unsaturation in organic compounds by electrometric titration. Analytical Chemistry 21: 1461-1465. Page .

exhibit 31 of chapter 7

Minutes required for 1 inch of gas space above a membrane to reach a relative humidity of 25%

A) The TIMES - in minutes

Thickness	Material	Times required
0.03	Cellulose acetate	0.77,0.83,0.57,0.53,0.62,
0.13	Cellulose acetate	3.92,5.03,4.78,4.47,5.35,
0.03	Cellophane	0.78,0.83,0.82,0.85,0.57,
0.05	Polyvinyl alcohol	1.65,1.55,1.75,1.83,1.08,
0.03	Nylon	1.75,1.40,1.25,1.25,1.65,
0.04	Nylon	3.67,3.60,3.53,3.63,3.32,
0.07	Vinylite plasticizer a	6.78,7125,7.25,7.67,5.10,
0.13	Vinylite plasticizer a	22.67,22.47,30.27,29.50,30.15,
0.11	Vinylite plasticizer b	16.13,16.23,15.78,16.23,17.20,
0.08	Vinylite plasticizer c	18.90,23.78,20.25,23.63,22.50,
0.13	Vinylite plasticizer d	30.06,35.12,39.42,42.67,42.33,
0.10	Vinylite plasticizer e	12.73,17.02,15.95,17.85,17.35,

B) The SOURCE

E. R. Weaver and R. Riley 1948. Measurement of water in gases by electrical conduction in a film of hygroscopic material. (Use of pressure changes in calibration) Analytical Chemistry 20: 216-229.

exhibit 32 of chapter 7

Solid phases of ethylenediamine and salts

A) The DATA

Salt	Mole ratios in solid phase: ethylemediamine/salt
NaI	3.0,3.0,3.3,2.9,3.1,
$NaClO_4$	2.9,3.3,3.3,3.2,3.0,3.0,
LiBr	2.0,2.6,2.1,1.9,2.0,
LiCl	2.0,1.9,2.0,
LiI	3.0,
$BaCl_2$	3.75,4.1,4.0,
$CaCl_2$	6.2,5.9,
$SrCl_2$	5.8,6.2,6.2,

B) The SOURCE

H. S. Isbin and K. A. Kobe 1945. The solubility of some salts in ethylenediamine, monoethanolamine and ethylene glycol. Journal of the American Chemical Society 67: 464-465. Table on page .

exhibit 33 of chapter 7

Electrical resistivity of fibers

A) The DATA

Material and relative humidity	Resistivity, 10^9 ohm-cm^2/cm
Nylon monofilament -340 denier, 64%	12,13,12,12,16,13,12,13,
Nylon monofilament -30 denier, 64%	32,26,26,29,29,41,32,
Nylon ton -3 denier, 85%	1.00,0.95,1.00,0.95,1.00,1.00,1.00, 0.95,1.10,0.90,1.00,
Human hair ∿0.001 diameter 85%	5.7,6.8,6.2,6.2,
Wool Columbia 58's, 85%	0.095,0.092,0.089,0.075,0.050,0.087,

B) The SOURCE

S. P. Hersh and D. J. Montgomery 1952. Electrical resistance measurements on fibers and fiber assemblies. Textile Research Journal 22: 805-818. Table III on page 810.